REMARKS

STATUS OF CLAIMS

Claims 1-34 were pending in the present application, with claims 1, 16, 17, 22, 25 and 27 being independent. Claims 1, 3, 16, 17, and 25-27 have been amended, claims 22-24 have been cancelled while no claims have been added. It should be noted that claim 3 has been amended to correct a spelling error and not for reasons related to patentability. Therefore, claims 1-21 and 25-34 are pending and are submitted for reconsideration.

CLAIM OBJECTIONS

The objection to claim 22 is moot in view of the proposed cancellation of claim 22.

REJECTIONS UNDER 35 U.S.C. §101

In the office action, claims 1, 16, 17, 22, and 27 have been rejected under 35 U.S.C. §101. In reply, applicants have amended claims 1, 16, 17, and 27 to recite hardware in the body of the claims as suggested in the office action while the rejection is moot with respect to cancelled claim 22. Accordingly, while applicants disagree with the legal basis of this rejection, applicants submit that the pending claims meet the requirements of section 101 as currently interpreted by the PTO.

REJECTION UNDER 35 U.S.C. §112

In the office action, claims 17, 22, and 27 are rejected under 35 U.S.C. §112 for allegedly failing to comply with the enablement requirement. The office action states that the recited feature "wherein the existing collection of data records does not comprise manually segmented training data," apparently contrasts with page 3, lines 2-6 of the specification. However, the specification very clearly states that the system uses a reference table or relationship but does not require labeled data (beyond what is present in a reference table or relationship). As clarified on page 3, lines 7-10, a representative system uses an existing reference table (or relation) having clean records or tuples that would typically be available in data warehouses for training purposes. Accordingly, it is clear that this recited feature uses existing reference tables (or relations) for training purposes and does not require any manual segmentation of the *training* data (i.e., data that is used in the training process is not manually segmented for training purposes). For

example, the model building component builds a number of attribute recognition models based on an existing relation of data records as recited in claim 17. Accordingly, applicants submit that these claims are sufficiently enabled in the specification so that one skilled in the art could make or use the claimed features without undue experimentation.

REJECTIONS UNDER 35 U.S.C. §103

Claims 1-34 are rejected under 35 U.S.C. 103(a) as being unpatentable over Borkar et al., "Automatic segmentation of text strings into structured records" (hereafter "Borkar") in view of Ando et al., "Mostly Unsupervised Statistical Segmentation of Japanese Sequences," (hereafter "Ando"). Claims 14, 20, and 23 are rejected under 35 U.S.C. 103(a) as being unpatentable over Borkar and Ando in view of U.S. Patent No. 5,095,432 ("Reed"). Applicants respectfully traverse this rejection for at least the following reasons.

Independent claim 1 recites, *inter alia*, (1) providing a state transition model based on an existing collection of data records that includes probabilities to segment input strings into component parts wherein the existing collection of data records (used to provide or train the state transition model) does not comprise manually segment training data, and (2) adjusting the probabilities (in the state transition model) to account for erroneous token placement in the input string. These recited features are not disclosed or suggested by the applied references.

With respect to feature (1), as noted previously, Borkar discloses a supervised model-based approach for text segmentation in which scalability is achieved by automatically learning segmentation models from manually tagged or segmented training data. An inherent limitation in supervised model-based approaches is that it is often difficult to obtain sufficient training data, especially data that is comprehensive enough to illustrate all features of test data (e.g., see Borkar, page 10, section 3.5, 1st paragraph). To cure this deficiency in Borkar (which is acknowledged in the office action), the office action cites to Ando and specifically to page 2, line 26-30 of Ando. However, Ando discloses a *mostly* unsupervised statistical segmentation and states that a "small number of pre-segmented training examples" are needed for good performance. See page 2, lines 31-32 of Ando. Therefore, even Ando requires at least a small number of pre-segmented

training examples in sharp contrast to recited feature (1) which does not require any manually segmented training data. Accordingly, neither of the applied references discloses recited feature (1) and nor does their reasonable combination. Accordingly, recited feature (1) is not disclosed or suggested by the applied references.

Feature (2) recites adjusting the probabilities (in the state transition model) to account for erroneous token placement in the input string. This recited feature provides the advantage of a robust segmentation system that can deal with input errors. For example, the specification teaches an attribute recognition model (ARM) topology that relaxes one or more of a positional specificity, sequential specificity, or token specificity to improve segmentation robustness to input errors. One method of relaxing positional specificity is by categorizing attribute values into three positions: beginning, middle, and trailing positions. See page 11, line 12 to page 12, line 31 of the specification. No such adjustment of probabilities to account for erroneous token placement in input strings is disclosed by the applied references. Specifically, the office action cites to page 7, section 2.5.1, lines 16-21 of Borkar with respect to this feature. However, the cited portions only disclose that the model is set up such that certain symbol-state assignments are invalid in determining the probability of the most probable path that generates a particular output sequence. However, this disclosure does not disclose adjusting probabilities that account for erroneous token placement in the input string. Accordingly, this recited feature is also not disclosed by any of the applied references.

Since several recited features are not disclosed by the applied references, the office action fails to make a *prima facie* case of obviousness with respect to independent claim 1. Likewise, independent claims 17 and 27 also recite features that are similar to features (1) and (2) discussed above. Accordingly, these independent claims are also patentable for reasons that are similar to that discussed above with respect to claim 1.

Independent claim 16 recites, *inter alia*, (1) providing a state model by analyzing substrings in a reference table of string records wherein the reference table of string records does not comprise manually segmented data; and (2) analyzing the substrings with the state model that provides a beginning, middle, and trailing token topology for each attribute. These features are also distinguishable over the applied combination of Borkar and Ando for reasons that are similar to the features (1) and (2) discussed earlier

with respect to claim 1. Specifically, neither of the references teach training of a state model by analyzing a reference table of string records without using any manually segmented data. Likewise, neither of the references discloses the specific feature of increasing robustness to errors in the input string data by analyzing the substrings within an attribute by assuming a beginning, middle, and a trailing topology with the topology including a token for an empty attribute component. Accordingly, these recited features are also disclosed by the applied references and claim 16 is also patentable over the applied references.

DEPENDENT CLAIMS

The dependent claims are patentable for at least the same reasons as the independent claims on which they ultimately depend. In addition, they recite additional features which are also patentable when considered as a whole.

For example, dependent claims 14, 20, and 23 recite additional features which are disclosed by the applied references. With respect to these features the office action acknowledges that these features are not disclosed by Borkar and Ando. However, the office action cites to Reed with respect to these features. However, Reed does not disclose these recited features either.

Specifically, claim 14 recites that the state transition model has a beginning, middle, and a trailing state topology (for each attribute) and the process of accounting for misordered or inserted tokens is performed by copying states from one of said beginning, middle, or trailing states into another of the said beginning, middle, or trailing states. No such 3 part state topology or processing for accounting for misordered or inserted tokens is disclosed by Reed. Specifically, the cited portion of Reed only discloses construction of state set of an RVG grammar by using three operations of predict, shift, and complete. See col. 4, lines 42-45. These operations may add more states to state Si being processed or add a new state Si+1. See col. 4, lines 44-46 of Reed. The cited portion col. 5, lines 1 discloses that a copy of state Si is added to state Si+1 with its vector updated using the specified rv operation. However, this building of the states does not in any way relate to the claimed 3 part state topology for each attribute in which the process of misordered or inserted tokens is accounted for by copying states from one of the beginning, middle, or trailing states to another of the said beginning, middle, or trailing states. Accordingly,

this recited feature is not disclosed by any of the applied references or their reasonable combination. Therefore, claim 14 is patentable for this additional reason. Likewise, claims 20 and 23 are also further patentable for reasons that are analogous to that discussed above with respect to claim 14.

CONCLUSION

Accordingly, in view of the above amendments and remarks it is believed that the application is now in condition for allowance. If the Examiner believes, after this amendment, that the application is not in condition for allowance, the Examiner is requested to call the applicants' attorney at the telephone number listed below.

If this response is not considered timely filed and if a request for an extension of time is otherwise absent, applicants hereby request any necessary extension of time. If there is a fee occasioned by this response, including an extension fee that is not covered by an enclosed check please charge any deficiency to Deposit Account No. 50-0463.

Respectfully submitted,

Microsoft Corporation

Date: August 29, 2007

By: /Aaron C. Chatterjee/

Aaron C. Chatterjee, Reg. No.: 41,398
Attorney for Applicants
Direct telephone: (703) 647-6572
Microsoft Corporation
One Microsoft Way
Redmond WA 98052-6399

CERTIFICATE OF MAILING OR TRANSMISSION [37 CFR 1.8(a)]

I hereby certify that this correspondence is being electronically deposited with the USPTO via EFS-Web on the date shown below:

August 29, 2007
Date

/Kate Marochkina/
Signature

Kate Marochkina
Type or Print Name